Journal of Air Pollution and Health (Winter 2025); 10(1): 61-82

Original Article

Available online at http://japh.tums.ac.ir

Ground data analysis for $PM_{2.5}$ Prediction using predictive modeling techniques

Elham Nourmohammad, Yousef Rashidi*

Department of Environmental Technologies, Shahid Beheshti University, Tehran, Iran

ARTICLE INFORMATION

Article Chronology:
Received 26 January 2025
Revised 22 February 2025
Accepted 01 March 2025
Published 29 March 2025

Keywords:
Particulate matters ($PM_{2.5}$) prediction; Ground data; Meteorological data; Traffic emissions; Air quality management

CORRESPONDING AUTHOR:

y_rashidi@sbu.ac.ir
Tel : (+98 21) 22432040
Fax : (+98 21) 22432040

ABSTRACT

Introduction: Air quality forecasting, particularly predicting Particulate Matter ($PM_{2.5}$) concentrations, has gained significant attention due to its critical implications for public health and environmental management. Accurately predicting $PM_{2.5}$, a harmful air pollutant associated with respiratory and cardiovascular diseases, is vital for effective air quality management in densely populated urban areas.

Materials and methods: This study uses various meteorological and environmental data combinations in Tehran, Iran, this study investigates the efficacy of three predictive modeling techniques Auto Regressive Integrated Moving Average (ARIMA), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) in forecasting daily and monthly $PM_{2.5}$ levels. The models were evaluated based on performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 scores.

Results: Results indicate that XGBoost excelled in daily predictions when using solely meteorological data, achieving an R^2 score of 0.998674, while ARIMA demonstrated strong predictive capacity but struggled with added complexity. LSTM maintained reasonable performance amidst increased data input but faced challenges in both daily and monthly forecasts. Monthly predictions from all models proved less reliable, particularly with ARIMA yielding negative R^2 values, indicating suboptimal performance compared to simplistic models.

Conclusion: The findings highlight the importance of model selection and feature engineering in accurately predicting $PM_{2.5}$ levels. The study suggests a shift towards hybrid modeling approaches and incorporating diverse environmental data to enhance forecasting accuracy in air quality management, particularly for long-term predictions.

Please cite this article as: Nourmohammad E, Rashidi Y. Ground data analysis for $PM_{2.5}$ Prediction using predictive modeling techniques. Journal of Air Pollution and Health. 2025;10(1): 61-82.

Introduction

Air pollution has emerged as a quintessential global environmental challenge, significantly impacting urban populations worldwide. Modern cities are increasingly burdened by deteriorating air quality, with alarming statistics indicating that approximately 92% of the world's population resides in areas exceeding the World Health Organization's (WHO) air quality guidelines [1-3]. The WHO and the International Agency for Research on Cancer (IARC) have classified air pollution as a human carcinogen, resulting in approximately 9 million deaths annually, accounting for 16% of global fatalities. If current trends persist, air pollution will become the leading cause of premature death by 2050 [4].

The detrimental health effects linked to air pollution are extensive, including aggravated cardiovascular and respiratory illnesses, asthma, and emphysema. Research indicates that ambient air pollution has shortened global lifespans by an average of 1.8 years. Furthermore, the economic ramifications are staggering, with an estimated global cost of $ 5 trillion each year due to premature deaths, healthcare expenses, and lost labor [5]. Among the various pollutants, Particulate Matter (PM), particularly $PM_{2.5}$ and PM_{10}, has garnered increased attention due to their significant health impacts and ecological consequences [6, 7] and respiratory ailments [8, 9]. Notably, exposure to $PM_{2.5}$ resulted in approximately 4.58 million deaths in 2017, with ambient $PM_{2.5}$ responsible for 64.2% of these fatalities [10]. Despite its known risks, the prediction and monitoring of $PM_{2.5}$ levels remain complex due to various factors, including the sparse availability of ground monitoring stations in urban areas and the challenges posed by urbanization, population density, and high operational costs [11].

Tehran, the capital of Iran, presents a unique case study in air quality management due to its geographical features, climatic conditions, and rapid urbanization. Nestled within a valley and surrounded by mountains, the city experiences meteorological phenomena that can trap pollutants close to the ground, exacerbating air pollution levels. Furthermore, high levels of vehicular traffic, industrial emissions, and construction activities contribute to the persistent presence of $PM_{2.5}$ in the air. Understanding the dynamics of $PM_{2.5}$ pollution in Tehran is essential for implementing effective public health interventions and environmental policies [12, 13].

Accurate forecasting of $PM_{2.5}$ concentrations is vital for timely decision-making and effective air quality management. Traditional approaches such as AutoRegressive Integrated Moving Average (ARIMA) have been widely used for time series forecasting [14, 15]; however, they often struggle to capture the complex relationships within environmental data. The rise of machine learning techniques, particularly XGBoost (Extreme Gradient Boosting) and Long Short-Term Memory (LSTM) networks, offers exciting new possibilities [16, 17]. These modern algorithms have shown great promise in handling large datasets, nonlinear correlations, and time-dependent patterns.

This study aims to explore and compare the predictive performance of ARIMA, XGBoost, and LSTM models in forecasting daily and monthly $PM_{2.5}$ concentrations in Tehran. It utilizes a comprehensive dataset that includes meteorological factors, traffic data, and cloudiness information. The research seeks to achieve the following objectives:

1. Investigate the effectiveness of different modeling techniques in predicting $PM_{2.5}$ levels and identify the strengths and limitations of each method.

2. Examine the influence of meteorological variables and traffic patterns on $PM_{2.5}$ forecasting.

By leveraging advanced modeling techniques, this research aims to enhance the understanding of $PM_{2.5}$ dynamics, ultimately aiding in the development of effective public health interventions and environmental policies aimed at reducing air pollution levels.

2. Theoretical modeling

This study evaluates the predictive performance of $PM_{2.5}$ concentration estimation across various configurations of ground and satellite data in Tehran. Specifically, we designed three distinct sections to analyze the effects of combining different datasets on the accuracy of $PM_{2.5}$ predictions.

Materials and methods

Study area

Tehran is located between latitude 35° 35′ N and 35° 48′ N and longitude 51° 17′ E to 51° 33′ E. The city is situated at an elevation exceeding 1,200 m above sea level, with a topographical range of 700 m between its highest and lowest points. This urban area is home to approximately 13.3 million residents, supplemented by around 10 million commuters [18]. Air quality in central Iran, particularly in Tehran, is sometimes compromised by dust storms originating from various sources [19]. However, the most significant contributors to air pollution are localized anthropogenic factors, including rapid demographic growth, and the conversion of agricultural lands and natural areas into urban spaces. Notably, mobile sources (vehicles) account for nearly 85% of total pollutants and 70% of PM emissions [20]. The surrounding Alborz Mountains to the north and Bibi Shahrbanoo Mountain to the southeast exacerbate pollution by channeling winds that carry pollutants from western industrial zones to the eastern parts of Tehran [21] (Fig. 1).

Fig. 1. Geographical location of Tehran province

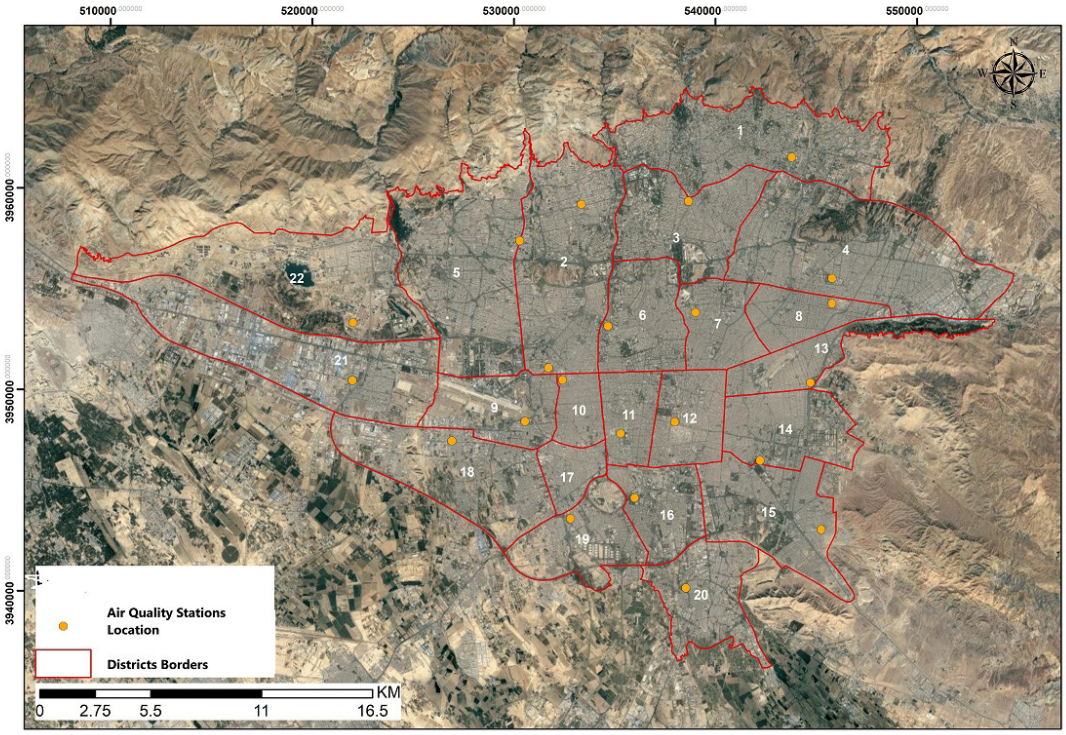


Fig. 2. Geographic locations of Tehran's air quality stations

Data collection

We collected data from 22 air quality monitoring stations across the Tehran metropolitan area. The dataset included meteorological variables (temperature, wind speed, humidity, precipitation, and boundary layer height) and $PM_{2.5}$ concentration levels, district-specific traffic, and cloudiness data. These localized observations are essential for understanding pollution dynamics within Tehran (Fig. 2). Meteorological data were sourced from the Iranian Meteorological Organization, traffic data from the Greater Tehran Traffic Control Company, and $PM_{2.5}$ measurements from the Air Quality Control Site.

Preprocessing and post processing of ground and traffic data

To ensure the reliability and integrity of the dataset, we employed advanced estimation methods to reconstruct incomplete time series for both ground and traffic data. This preprocessing phase was crucial in preparing the data for subsequent analysis and modeling.

Manging missing dadat

Ground data interpolation

For missing values in meteorological measurements, we utilized spatiotemporal interpolation techniques such as kriging. This method leverages both spatial and temporal correlations within the data, which enables more accurate estimates of missing values by considering the relationships between nearby observations. For continuous variables with smaller gaps, linear interpolation was applied, estimating missing values based on adjacent time points. This approach was particularly effective for variables like $PM_{2.5}$ concentrations and meteorological data exhibiting stable trends over time, ensuring that minor gaps did not

compromise the dataset's integrity.

Advanced compensation techniques

In scenarios where variables exhibited low variance, such as temperature and wind speed, we implemented mean or median compensation methods. These measures of central tendency provided reliable estimates for the missing data points, thereby allowing the dataset to remain robust and representative of actual conditions.

Missing date management in traffic data

The traffic dataset required specific management to achieve uniform temporal coverage, ensuring that there were exactly 365 days of data for each monitoring station. Missing dates were systematically added, and interpolated values for those dates were derived from adjacent observations. This process was crucial for capturing the seasonal effects of traffic levels on $PM_{2.5}$ concentrations, providing essential data for subsequent analyses.

Traffic data interpolation

In light of inconsistencies in traffic data stemming from varying reporting standards and monitoring infrastructure across regions, we implemented several strategies to ensure consistency. For variables with complex relationships, such as traffic congestion, we applied K-Nearest Neighbors (KNN) compensation. This technique estimates missing values by considering the nearest neighbors in the dataset, thereby preserving the relationships and patterns inherent in the traffic data.

For filling temporal gaps in traffic data, we relied on historical trends and nearest-neighbor methods, which helped maintain continuity in the time series. Furthermore, outliers were detected and removed using thresholds defined by methodologies such as the Interquartile Range (IQR) and Z-score thresholds. These steps enhanced the integrity of the dataset while minimizing noise, ensuring that extreme values did not skew the results.

Spatial coordination

Traffic data were spatially linked to specific areas of Tehran using geographic coordinates that corresponded with the locations of air quality monitoring stations. This spatial mapping enabled an analysis of how traffic impacts $PM_{2.5}$ concentrations at a regional level, facilitating insights into localized pollution sources and allowing for targeted interventions.

Temporal coordination

To ensure compatibility across all datasets, traffic data were collected and aligned with the temporal resolution of other datasets, including satellite and atmospheric data. This synchronization ensured consistency in the modeling process and allowed for a coherent analysis of relationships among the various factors influencing $PM_{2.5}$ levels.

To ensure compatibility across all datasets, traffic data were collected and aligned with the temporal resolution of other datasets, including atmospheric data. This synchronization ensured consistency in the modeling process, enabling coherent analyses of the relationships among the various factors influencing $PM_{2.5}$ levels.

Normalization and scaling

Normalization was necessary to eliminate unit differences among the various features and to ensure a balanced contribution to the machine learning models. We applied Min-Max scaling to criteria such as traffic density and flow, transforming the values to a standard range (e.g., 0 to 1). This step was critical in preventing any feature from disproportionately influencing model outcomes, thereby enhancing overall model performance and stability [22-24].

Air quality data and meteorological data handling and validatio

In addition to the aforementioned techniques, we paid careful attention to the handling and validation of air quality and meteorological data. The air quality data, primarily $PM_{2.5}$ concentrations, were subjected to rigorous quality control. This involved identifying and correcting anomalies, such as spikes caused by measurement errors or equipment malfunctions. Sensor calibration and periodic validation against reference data ensured the accuracy of these measurements.

Meteorological data were similarly validated through comparisons with established weather stations in the vicinity. Any discrepancies were addressed through adjustment procedures based on reliable historical data patterns. We implemented cross-validation techniques to assess the reliability of the data, allowing us to evaluate the consistency and robustness of the estimates derived from different processing techniques.

Comprehensive documentation of the preprocessing steps, including imputation and normalization methods, was maintained to ensure reproducibility

Prediction models

We structured the dataset to support daily and monthly $PM_{2.5}$ predictions, allowing us to apply machine learning, deep learning, and statistical models effectively.

Model selection and justification

We explored three modeling approaches to capture $PM_{2.5}$ dynamics:

Statistical model: ARIMA/SARIMA

ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) models are effective for time series forecasting, capturing linear trends and seasonality in stable datasets ARIMA models the non-seasonal Auto Regressive (AR) and Moving Average (MA) components, making it adept at handling linear trends. SARIMA extends ARIMA by incorporating seasonal components, which is especially useful for environmental datasets exhibiting regular seasonal patterns.

Machine learning model: XGBoost

We selected XGBoost for its computational efficiency and robust performance as an intermediate model.

Deep learning model: LSTM

LSTM architecture excels at modeling sequential data, making it suitable for capturing the temporal dependencies in air quality datasets.

Evaluation metrics

To assess the performance of the models, we used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2) (Eq. 1). These metrics allowed for a comprehensive comparison of each model's effectiveness in predicting daily and monthly $PM_{2.5}$ levels.

$$RMSE = \sqrt{((\sum_{i=1}^{n}(y_i - (y_i)\hat{}))/n)}$$

$$R^2=1- (\sum_{i=1}^{n}(y\hat{}_i - y_i)^2)/\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad (1)$$

$$MAE= 1/n \sum_{i=1}^{n} [\![| y_i - y\hat{}_i |]\!]$$

where:

n=number of observations

y_i=the actual value of i^th observations

y^_i=the predicted value of i^th observation

y¯_i=the mean value of all observations

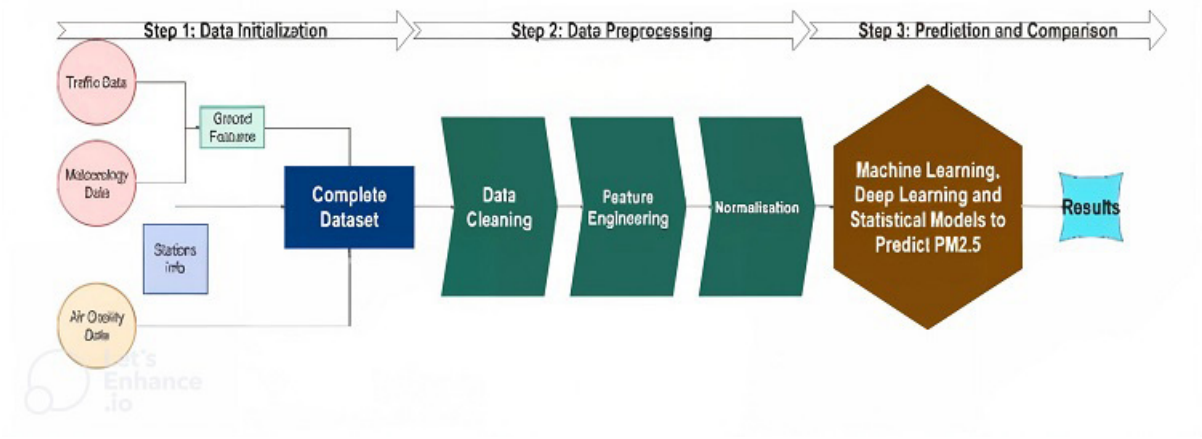


Fig. 3. Data preprocessing workflow

Results and discussion

This study evaluated the predictive performance of SARIMA/ARIMA, XGBoost, and LSTM models across various configurations for predicting $PM_{2.5}$ levels. The primary focus was on daily $PM_{2.5}$ predictions, essential for short-term air quality management, while monthly analyses provided supplementary insights into long-term trends. Key performance metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) were utilized to assess the accuracy and robustness of each model.

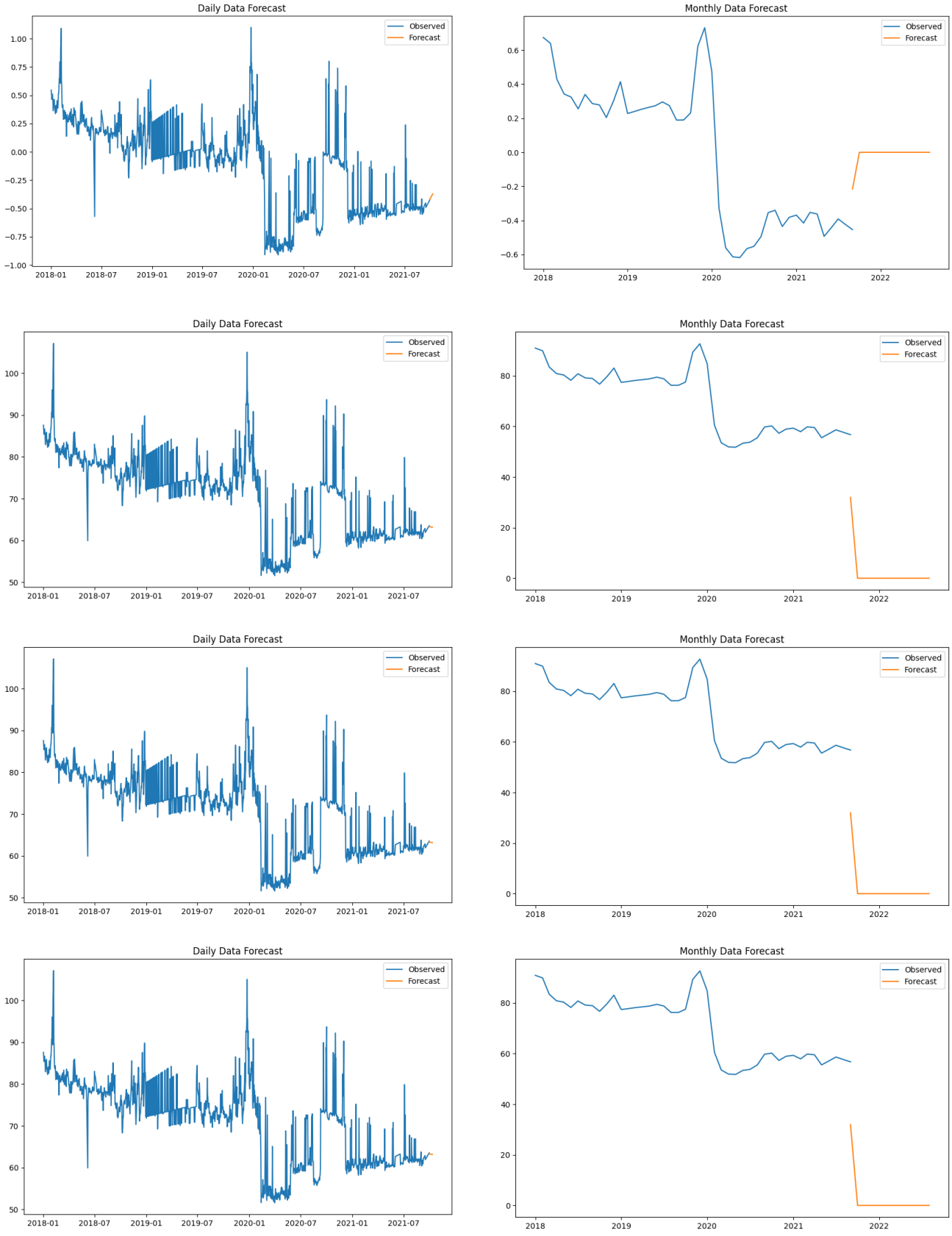
Model performance overview

Meteorological data

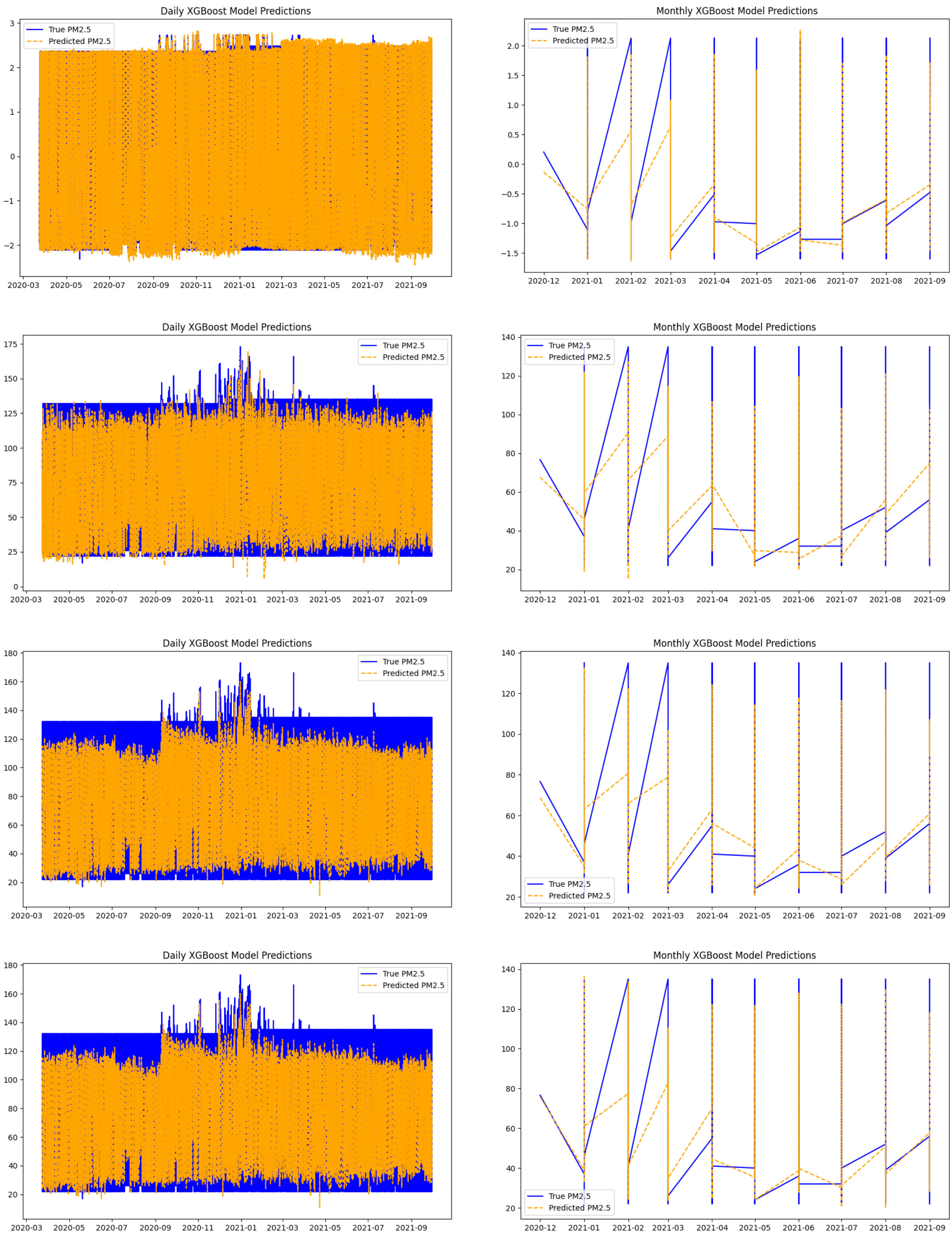
ARIMA when using meteorological data alone, indicates a high level of predictive accuracy (R^2 = 0.934968). XGBoost achieved the best R^2 score (0.998674), showcasing its robustness in capturing patterns in the data (despite higher RMSE and MAE compared to ARIMA). LSTM produced a less favorable performance than ARIMA and XGBoost, suggesting challenges in accurately forecasting $PM_{2.5}$ levels with this model when limited to meteorological data (Fig. 4).

Impact of cloudiness and traffic data

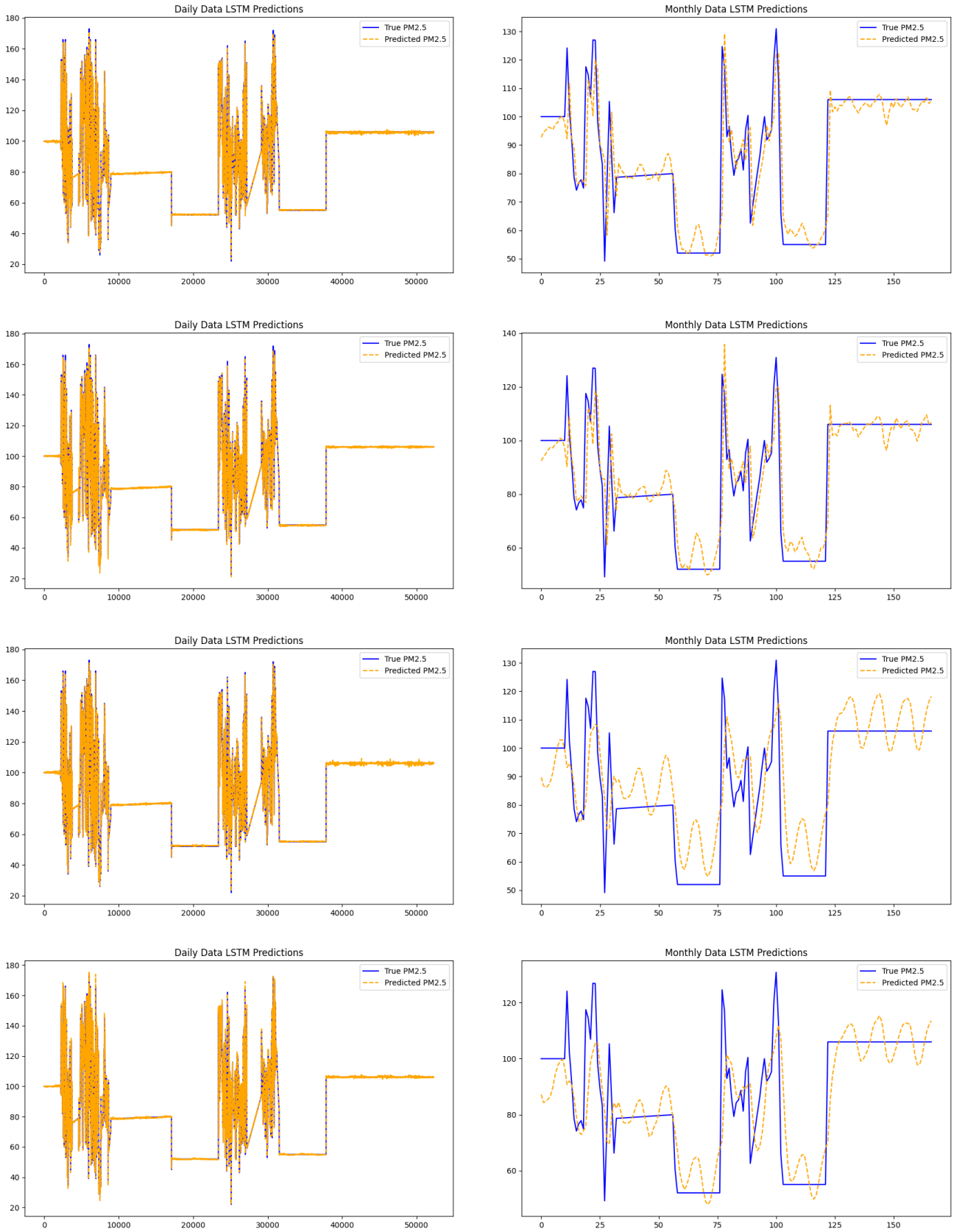
When additional features, such as cloudiness and traffic data, were included, the performance of ARIMA and XGBoost significantly deteriorated. In contrast, LSTM showed relative stability, maintaining a level of performance similar to that of the configuration using only meteorological data. This indicates that LSTM is robust and capable of handling more complex data configurations without significantly declining prediction quality (Fig. 4).



A) Arima model



B) XGBoost model



C) LSTM model

Fig. 4. Prediction $PM_{2.5}$ using meteorology data and other features using Arima, XGBosst, and LSTM

Daily predictions

Table 1 summarizes the performance metrics of the ARIMA, XGBoost, and LSTM models for daily $PM_{2.5}$ predictions across different data configurations, including meteorological data, cloudiness, and traffic data.

Meteorological data

ARIMA demonstrated exceptionally low RMSE (0.004147) and MAE (0.003020), indicating high predictive accuracy (R^2 = 0.934968). XGBoost achieved the best R^2 score (0.998674) with the same dataset, showcasing strong performance in capturing complex patterns despite higher RMSE (0.042189) and MAE (0.024427).LSTM exhibited less favorable performance with significantly higher RMSE (2.931065) and MAE (0.624375), highlighting challenges in accurately forecasting $PM_{2.5}$ levels with solely meteorological data.

Impact of cloudiness and traffic data

When incorporating cloudiness and traffic data, Both ARIMA and XGBoost displayed considerable deterioration in performance when additional features were incorporated. ARIMA's RMSE substantially increased to 0.269315, and R^2 dropped to 0.546620, suggesting significant challenges with multi-dimensional datasets. XGBoost's RMSE surged to 12.934760, with a decrease in R^2 to 0.795519, indicating reduced predictive power. In contrast, LSTM maintained a relatively stable performance with RMSE values around 2.9 suggesting its robustness in handling complex data configurations without significant degradation in prediction quality.

Table 1. Model performance on daily $PM_{2.5}$ prediction across different data configurations

Model Configuration	Dataset	Frequency	RMSE	MAE	R^2
ARIMA	Meteorological Data Only	Daily	0.004147	0.003020	0.934968
XGBoost	Meteorological Data Only	Daily	0.042189	0.024427	0.998674
LSTM	Meteorological Data Only	Daily	2.931065	0.624375	0.986079
ARIMA	Meteorological Data + Cloudiness	Daily	0.269315	0.260328	0.546620
XGBoost	Meteorological Data + Cloudiness	Daily	12.934760	9.933494	0.795519
LSTM	Meteorological Data + Cloudiness	Daily	2.910194	0.587228	0.986276

Table 1. Continued

Model Configuration	Dataset	Frequency	RMSE	MAE	R^2
ARIMA	Meteorological Data + Traffic Data	Daily	0.269303	0.260312	0.546662
XGBoost	Meteorological Data + Traffic Data	Daily	13.459413	10.183478	0.778645
LSTM	Meteorological Data + Traffic Data	Daily	2.918040	0.603696	0.986202
ARIMA	Meteorological Data + Cloudiness + Traffic Data	Daily	0.269303	0.260312	0.546662
XGBoost	Meteorological Data + Cloudiness + Traffic Data	Daily	13.459413	10.183478	0.778645
LSTM	Meteorological Data + Cloudiness + Traffic Data	Daily	2.886912	0.586673	0.986495

Monthly data

Table 2 outlines the prediction performance for monthly $PM_{2.5}$ levels. Notably, all models faced significant challenges when including cloudiness and traffic data.

Meteorological data

The ARIMA model showed poor performance, producing negative R^2 values (-25.019984), indicating that using the mean for predictions was more effective. Its RMSE and MAE were relatively high compared to XGBoost and LSTM. XGBoost and LSTM provided low R^2 values (0.767700 and 0.772545, respectively), underscoring difficulties in accurately forecasting monthly averages.

Impact of cloudiness and traffic data

The inclusion of additional features severely impacted all models, with ARIMA yielding a negative R^2 value of -481.007507, indicating poor fit. XGBoost, while better than ARIMA, experienced declining predictive accuracy, demonstrated by its rising RMSE and reducing R^2 values.LSTM also struggled with high RMSE values in monthly predictions, indicating that despite its effectiveness in daily forecasts, it had difficulty generalizing across extended time frames.

Table 2. Model performance on monthly $PM_{2.5}$ prediction across different scenarios

Model Configuration	Dataset	Frequency	RMSE	MAE	R^2
ARIMA	Meteorological Data Only	Monthly	0.227687	0.223799	-25.019984
XGBoost	Meteorological Data Only	Monthly	0.538045	0.406533	0.767700
LSTM	Meteorological Data Only	Monthly	10.510049	6.375597	0.772545
ARIMA	Meteorological Data + Cloudiness	Monthly	29.723305	29.261299	-481.007507
XGBoost	Meteorological Data + Cloudiness	Monthly	16.523732	12.568438	0.761846
LSTM	Meteorological Data + Cloudiness	Monthly	10.395830	6.135935	0.777462
ARIMA	Meteorological Data + Traffic Data	Monthly	29.723305	29.261299	-481.007507
XGBoost	Meteorological Data + Traffic Data	Monthly	17.119665	11.915684	0.744358
LSTM	Meteorological Data + Traffic Data	Monthly	10.881068	6.628909	0.756203
ARIMA	Meteorological Data + Cloudiness + Traffic Data	Monthly	29.723305	29.261299	-481.007507
XGBoost	Meteorological Data + Cloudiness + Traffic Data	Monthly	18.824991	13.902366	0.690892
LSTM	Meteorological Data + Cloudiness + Traffic Data	Monthly	12.415167	8.579501	0.682611

Comparative insights between daily and monthly predictions

A distinct dichotomy exists between daily and monthly prediction performances. Daily models benefited from more straightforward data interactions, allowing for effective utilization of immediate atmospheric conditions and emissions. In contrast, monthly models faced increased complexity due to multicollinearity and potential overfitting, which impeded their predictive accuracy. Moreover, the smoothing effect associated with monthly averaging diminishes the granularity of data, leading to a loss of critical short-term dynamics essential for understanding pollution patterns [25, 26].

Model performance

We used the Mann-Whitney method to compare between models. The results of the Mann-Whitney U test and the corresponding Cohen's d effect sizes for pairwise comparisons between the models (ARIMA, XGBoost, and LSTM) are summarized below:

Root mean square error (RMSE)

ARIMA vs XGBoost: p-value = 0.369, Cohen's d = -1.97 (Large effect size). Although the p-value suggests no statistically significant difference, the large negative effect size indicates that XGBoost performs substantially better than ARIMA in terms of RMSE.

ARIMA vs LSTM

p-value = 0.077, Cohen's d = -30.57 (Extremely large effect size). This comparison indicates that LSTM significantly underperforms compared to ARIMA for RMSE, as supported by an extremely large effect size.

XGBoost vs LSTM: p-value = 0.658, Cohen's d = 1.36 (Moderate effect size). While there is no statistical significance, the moderate effect size suggests that XGBoost outperforms LSTM.

Mean absolute error (MAE)

ARIMA vs XGBoost: p-value = 0.4, Cohen's d = -1.95 (Large effect size). Similar to RMSE, XGBoost demonstrates a substantial advantage over ARIMA, although this p-value does not indicate statistical significance.

ARIMA vs LSTM: p-value = 0.1, Cohen's d = -4.98 (Extremely large effect size). This suggests that LSTM performs significantly worse than ARIMA for MAE, as evidenced by a considerable effect size.

XGBoost vs LSTM: p-value = 0.7, Cohen's d = 1.83 (Moderate effect size). As in previous comparisons, while not statistically significant, this finding indicates that XGBoost generally outperforms LSTM.

Coefficient of determination (R^2)

ARIMA vs XGBoost: p-value = 0.369, Cohen's d = -1.18 (Moderate effect size). XGBoost shows a moderate improvement over ARIMA, yet the p-value suggests that this difference is not statistically significant.

ARIMA vs LSTM: p-value = 0.077, Cohen's d = -2.40 (Large effect size). LSTM's performance is notably weaker than that of ARIMA, supported by a large effect size.

XGBoost vs LSTM: p-value = 0.658, Cohen's d = -1.83 (Moderate effect size). Despite the lack of statistical significance, XGBoost's performance is moderately better than LSTM.

The analysis reveals that XGBoost generally exhibits substantial advantages over ARIMA, particularly in terms of RMSE and MAE, although the p-values did not reach the

conventional threshold of 0.05 for statistical significance. Caution should be exercised when interpreting these results as definitive. LSTM consistently underperforms against both ARIMA and XGBoost, as evidenced by very large negative effect sizes in most of the comparisons.

Baseline statistical models

In this study, SARIMA/ARIMA served as baseline statistical models, providing a foundational benchmark for comparing various modeling approaches. While these models offer interpretability and reliability, they often fall short in capturing the complex non-linear interactions inherent in environmental data. This limitation underscores the necessity for employing advanced machine learning and deep learning techniques that can enhance air quality forecasting capabilities.

Advanced modeling

Both LSTM and XGBoost emerged as the most reliable models, effectively managing the complexities of high-dimensional and hybrid datasets. This finding aligns with existing literature, which shows that deep learning and ensemble-based methods consistently outperform traditional models in handling complex air quality datasets [27, 28].

As an ensemble-based tree method, XGBoost excels at capturing intricate interactions among input variables, demonstrating robust performance in environmental modeling tasks. Its ability to process large-scale datasets not only ensures computational efficiency but also enhances predictive power [29].

In our study, the risk of overfitting was a significant consideration, particularly concerning the XGBoost model, which demonstrated exceptionally high performance in daily $PM_{2.5}$ predictions. To address this risk, we implemented several measures throughout the modeling process. Firstly, we employed K-fold cross-validation techniques to assess model performance more reliably.This approach not only provides a more accurate estimate of model performance but also ensures thorough evaluation of the model's ability to generalize.

Secondly, we incorporated regularization techniques within the XGBoost framework. XGBoost offers parameters for preventing overfitting, such as L1 (Lasso regression) and L2 (Ridge regression) regularization. By tuning these parameters, we penalized model complexity, discouraging it from becoming overly tailored to the training data.

We also employed early stopping criteria during model training. By monitoring model performance on a validation dataset at each iteration, training can be halted if performance begins to degrade, thereby preventing overfitting to noise in the training data. This strategy is particularly effective for gradient boosting algorithms like XGBoost, which can quickly overfit if allowed to train for too many iterations.

A critical aspect of our approach involved validating the models on independent datasets. Whenever possible, we tested the models on separate validation datasets not used during training. This independent validation confirms that results are not merely artifacts of overfitting the training data. For instance, we conducted temporal validation by splitting the data into training and testing periods based on time, simulating real-world scenarios where future predictions rely on past data.

Additionally, we examined model performance across various configurations and assessed the spread of performance metrics to identify potential signs of overfitting. By comparing model performance between training and test

sets, we looked for significant discrepancies indicative of overfitting.

Daily forecasting models

The evaluation of daily $PM_{2.5}$ predictions revealed that ARIMA demonstrated commendable accuracy when relying solely on meteorological data. This ability suggests that ARIMA can effectively capture the linear relationships present in environmental data. However, XGBoost outperformed all models, showcasing its strength in modeling complex, non-linear interactions despite exhibiting higher RMSE and MAE values. The higher RMSE associated with XGBoost indicates that while the model can capture intricate patterns, it is also more sensitive to outliers and variability in the data.

Models like XGBoost are particularly well-suited for real-time daily forecasting, demonstrating strong performance under varying conditions. Their adaptability to changing data environments facilitates effective pattern recognition, enabling timely interventions in air quality management. The gradient-boosting framework of XGBoost combines multiple weak learners into a robust predictive model, enhancing accuracy and minimizing overfitting when properly tuned [30]. Previous studies have corroborated that XGBoost excels in daily $PM_{2.5}$ forecasting due to its ability to manage non-linear relationships effectively [30, 31].

In contrast to its strong performance in daily forecasting, LSTM exhibited limited effectiveness when reliant solely on meteorological data. This is evident from its higher RMSE, likely stemming from LSTM's dependence on historical sequential patterns, which may not be adequately captured with a smaller set of input features. These challenges underline the limitations that arise for LSTM when constrained to linear predictors in forecasting $PM_{2.5}$ levels.

Monthly forecasting models

All models encountered challenges in monthly predictions, particularly in configurations that combined multiple features. While XGBoost excelled in daily predictions, its performance declined significantly when applied to monthly data, especially with combined features. The negative R^2 values observed in ARIMA's monthly predictions highlight significant inadequacies within this model raising critical questions about its applicability for long-term forecasts in this context.

Several factors may contribute to ARIMA's poor performance. Firstly, ARIMA and SARIMA models rely on the assumptions of stationarity in time series data. The non-stationarity of $PM_{2.5}$ data—characterized by trends and seasonal fluctuations—can lead to the model fitting poorly. If the time series exhibits strong seasonality or trends that are not adequately addressed, the forecasts may be biased or ineffective.

Secondly, ARIMA models primarily capture linear relationships in data. The environmental factors influencing $PM_{2.5}$ levels often involve complex, non-linear interactions among variables (e.g., meteorological conditions, traffic volumes, and geographical influences). Thus, the linear nature of ARIMA may struggle to adequately model these intricate relationships.

Additionally, the effects of data aggregation must be considered. Monthly averaging of $PM_{2.5}$ data could diminish granularity, causing important short-term variations and events (like pollution spikes) to be masked. This averaging smooths out fluctuations, potentially leading to inaccuracies in capturing the underlying dynamics when using just meteorological data for prediction.

Moreover, the exclusion of significant predictive

variables, such as traffic emissions and seasonal indicators, from the ARIMA model adversely affects its predictive accuracy. These factors are particularly impactful over extended time frames and can drive variations in $PM_{2.5}$ concentrations.

To improve the situation, several strategies could be employed. Utilizing seasonal differencing or transformations before fitting the ARIMA model can help in removing seasonal effects. Enhancing the model structure to incorporate seasonal patterns with SARIMA can be beneficial.

Exploring hybrid models that combine ARIMA with machine learning or advanced statistical methods could potentially improve predictions. For example, a hybrid model could leverage ARIMA for capturing trends and seasonality, while utilizing machine learning techniques for capturing non-linear relationships (e.g., ARIMA + XGBoost).

Incorporating additional relevant predictors beyond meteorological data is equally crucial. Future modeling efforts should consider factors like traffic patterns, economic indicators, industrial emissions, or indices representing societal behaviors that impact pollution levels.

Moreover, evaluating alternative models can lead to better outcomes. Generalized Additive Models (GAMs) can effectively handle non-linear relationships and include multiple predictors while maintaining interpretability. Random Forest or Gradient Boosting methods can manage high-dimensional data and capture non-linearities effectively while reducing the risk of overfitting through systematic techniques.

Also, for monthly forecasting, alternative modeling approaches should be adopted, focusing on capturing longer-term trends while avoiding excessive variable inclusion. Techniques such as dimensionality reduction and feature selection are essential for enhancing predictive performance by minimizing multicollinearity among variables. Research has proposed hybrid models, such as the Weighted LSTM Neural Network (WLSTME), which integrate auxiliary meteorological data with historical $PM_{2.5}$ concentrations to effectively simulate spatial dependencies alongside temporal factors [32].

Limitations of models

The introduction of additional features, such as cloudiness and traffic data, led to significant performance declines for both ARIMA and XGBoost. This suggests that these models struggled to manage the increased noise in the data without sacrificing predictive power. For ARIMA, this deterioration indicates a loss of predictive accuracy when dealing with multi-dimensional datasets. In contrast, while LSTM networks exhibited consistent performance across various configurations, their efficacy diminished in monthly predictions. The temporal nature of meteorological data requires careful consideration of time lags and seasonal effects, which complicates the learning process for LSTM models, especially in non-stationary time series contexts [33].

However, LSTM demonstrated remarkable resilience by maintaining stable performance even with the added complexity, highlighting its robustness in scenarios involving multi-dimensional datasets—an area where traditional linear models may falter. For example, researchers showed that cloudiness can correlate with humidity and other weather-related variables, complicating analyses and potentially degrading prediction accuracy [29].

The introduction of additional features such as traffic and cloudiness data into the predictive models, specifically XGBoost and LSTM, has significant implications for model performance. While these features are essential for capturing the multifaceted nature of $PM_{2.5}$ dynamics, their

complexity can also impose challenges.

The complexity added by traffic data significantly affects both the XGBoost and LSTM models. Traffic patterns directly correlate with pollutant emissions, as vehicle movement contributes substantially to $PM_{2.5}$ levels in urban environments. However, the inclusion of traffic volume data can introduce collinearity among predictors, particularly if traffic data correlates with other meteorological variables such as wind speed and humidity. This collinearity may lead to model instability and can distort the understanding of the individual feature's contribution to the predictions.

Moreover, traffic data often have daily patterns that may not align well with the temporal resolutions used in the models. For instance, while certain features may fluctuate throughout the day (e.g., traffic volumes), the model may not effectively capture these dynamics if the temporal resolution is inconsistent. This can lead to a misrepresentation of the relationship between traffic volumes and $PM_{2.5}$ concentrations, ultimately diminishing predictive performance.

Similarly, cloudiness data adds complexity due to its potential nonlinear impacts. Cloud cover influences both temperature and solar radiation, which in turn can affect $PM_{2.5}$ dispersion and formation. However, the relationship is complex; for example, increased cloudiness can limit solar radiation, potentially leading to increased $PM_{2.5}$ accumulation in certain conditions, while also possibly reducing it under different meteorological contexts. These interactions may not be adequately captured by the models, leading to potential inaccuracies in predictions.

To mitigate the negative impacts of these complexities, several strategies can be employed. Firstly, conducting Exploratory Data Analysis (EDA) to understand the relationships among features before model fitting is crucial. Correlation matrices and scatter plots can help identify potential collinearity and interactions.

Future research directions

The divergence in model performance between daily and monthly predictions emphasizes the complexities inherent in time series forecasting for environmental data. Recognizing the intricate data interactions and employing appropriate modeling strategies are vital for capturing underlying patterns necessary for accurate forecasts.

Future research should explore hybrid models that combine strengths from various techniques, such as using XGBoost for feature extraction followed by LSTM for time series predictions. Additionally, expanding datasets to include more environmental and anthropogenic variables may enhance predictive accuracy and robustness, especially in monthly forecasting contexts. Integrating traffic data to account for daily variability while ensuring a greater focus on meteorological variables for seasonal modeling is also a promising avenue for advancement. The dynamics of $PM_{2.5}$ levels are influenced by a multitude of external factors beyond meteorological conditions. A comprehensive understanding of air quality requires consideration of economic activities and seasonal variations, as these aspects are integral to the variability of $PM_{2.5}$ concentrations in urban environments.

Economic activities play a significant role in air pollution levels. For instance, industrial emissions contribute notably to $PM_{2.5}$ concentrations. The type and scale of industrial operations in Tehran, including manufacturing, construction, and energy production, directly affect the volume of particulate matter released into the atmosphere. Additionally, economic growth and increased consumer activity lead to higher levels of vehicular traffic, which is another major source of $PM_{2.5}$. Elevated

traffic during peak hours can exacerbate air quality issues, particularly in urban centers. Therefore, incorporating economic indicators such as industrial output, traffic volume, and fuel consumption into predictive models could improve their accuracy and relevance.

Seasonal variations also play a critical role in shaping $PM_{2.5}$ levels. Different seasons in Tehran result in varying meteorological conditions that influence air quality. For example, temperature inversions during the winter months can trap pollutants close to the ground, leading to higher $PM_{2.5}$ concentrations than during warmer seasons when atmospheric mixing is more vigorous. Furthermore, certain environmental phenomena, such as dust storms prevalent in spring and summer, can lead to spikes in $PM_{2.5}$ levels that are not adequately captured by meteorological data alone. Future research should consider the impact of these two key factors in their models.

Conclusion

The observed dichotomy in model performance between daily and monthly $PM_{2.5}$ predictions emphasizes the critical importance of understanding data characteristics and dynamics. This study highlights that adapting modeling approaches based on these differences is essential for improving forecasting accuracy, which in turn enhances decision-making processes related to air quality management and mitigates health risks associated with $PM_{2.5}$ exposure. Analysis results reveal notable strengths and weaknesses across the three models evaluated. XGBoost emerged as the superior model for daily predictions with meteorological data, effectively capturing short-term variability. However, the significant decline in predictive performance when additional features were included underscores the necessity of careful feature selection for achieving accurate modeling. In contrast, while LSTM demonstrated robustness across varying configurations, its challenges with monthly predictions illustrate limitations in long-term forecasting unless underpinned by comprehensive data strategies. Moreover, the negative R^2 values obtained from ARIMA during monthly predictions point to the necessity for more sophisticated approaches that can address both the temporal complexities inherent in $PM_{2.5}$ data and the intricate interactions among various influencing factors.

To advance this field, future research should explore the integration of hybrid models that amalgamate strengths from various methodologies, thereby enhancing accuracy and mitigating overfitting risks. Additionally, focused efforts on expanding datasets to include a wider range of environmental variables could further bolster model robustness and elevate forecasting capabilities within the context of air quality prediction.

Financial supports

The present research did not receive any financial support.

Competing interests

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Acknowledgements

We would like to express our gratitude to all the organizations that assisted us in gathering information.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely ob-served by the authors.

References

1. Choi J, Oh JY, Lee YS, Min KH, Hur GY, Lee SY, et al. Harmful impact of air pollution on severe acute exacerbation of chronic obstructive pulmonary disease: particulate matter is hazardous. International journal of chronic obstructive pulmonary disease. 2018:1053-9.

2. Nabizadeh R, Yousefian F, Moghadam VK, Hadei M. Characteristics of cohort studies of long-term exposure to $PM_{2.5}$: a systematic review. Environmental science and pollution research. 2019;26:30755-71.

3. Hadei M, Naddafi K. Cardiovascular effects of airborne particulate matter: a review of rodent model studies. Chemosphere. 2020;242:125204.

4. Landrigan PJ, Fuller R, Acosta NJ, Adeyi O, Arnold R, Baldé AB, et al. The Lancet Commission on pollution and health. The lancet. 2018;391(10119):462-512.

5. Li T, Shen H, Yuan Q, Zhang X, Zhang L. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophysical Research Letters. 2017;44(23):11,985-11,93.

6. Bodor K, Szép R, Bodor Z. The human health risk assessment of particulate air pollution (PM2. 5 and PM_{10}) in Romania. Toxicology Reports. 2022;9:556-62.

7. Faraji Ghasemi F, Dobaradaran S, Saeedi R, Nabipour I, Nazmara S, Ranjbar Vakil Abadi D, et al. Levels and ecological and health risk assessment of $PM_{2.5}$-bound heavy metals in the northern part of the Persian Gulf. Environmental Science and Pollution Research. 2020;27:5305-13.

8. Cao S, Guo Q, Xue T, Wang B, Wang L, Duan X, et al. Long-term exposure to ambient $PM_{2.5}$ increase obesity risk in Chinese adults: A cross-sectional study based on a nationwide survey in China. Science of the Total Environment. 2021;778:145812.

9. Wu J, Zheng H, Zhe F, Xie W, Song J. Study on the relationship between urbanization and fine particulate matter ($PM_{2.5}$) concentration and its implication in China. Journal of cleaner production. 2018;182:872-82.

10. Bu X, Xie Z, Liu J, Wei L, Wang X, Chen M, et al. Global $PM_{2.5}$-attributable health burden from 1990 to 2017: Estimates from the Global Burden of disease study 2017. Environmental Research. 2021;197:111123.

11. Kim S-Y, Olives C, Sheppard L, Sampson PD, Larson TV, Keller JP, et al. Historical prediction modeling approach for estimating long-term concentrations of $PM_{2.5}$ in cohort studies before the 1999 implementation of widespread monitoring. Environmental health perspectives. 2017;125(1):38-46.

12. Naddafi K, Hassanvand MS, Yunesian M, Momeniha F, Nabizadeh R, Faridi S, et al. Health impact assessment of air pollution in megacity of Tehran, Iran. Iranian journal of environmental health science & engineering. 2012;9:1-7.

13. Pardakhti A, Baheeraei H, Dehhaghi S. Forecasting and Seasonal Investigation of PM_{10} Concentration Trend: a Time Series and Trend Analysis Study in Tehran. Pollution. 2023;9(4):1579-88.

14. Gao W, Xiao T, Zou L, Li H, Gu S. Analysis and Prediction of Atmospheric Environmental Quality Based on the Autoregressive Integrated

Moving Average Model (ARIMA Model) in Hunan Province, China. Sustainability. 2024;16(19):8471.

15. Ramadan MS, Abuelgasim A, Al Hosani N. Advancing air quality forecasting in Abu Dhabi, UAE using time series models. Frontiers in Environmental Science. 2024;12:1393878.

16. Zaini Na, Ean LW, Ahmed AN, Abdul Malek M, Chow MF. $PM_{2.5}$ forecasting for an urban area based on deep learning and decomposition method. Scientific Reports. 2022;12(1):17565.

17. Saminathan S, Malathy C. Ensemble-based classification approach for $PM_{2.5}$ concentration forecasting using meteorological data. Frontiers in big Data. 2023;6:1175259.

18. Habibi R, Alesheikh AA, Mohammadinia A, Sharif M. An assessment of spatial pattern characterization of air pollution: A case study of CO and $PM_{2.5}$ in Tehran, Iran. ISPRS international journal of Geo-information. 2017;6(9):270.

19. Arhami M, Hosseini V, Shahne MZ, Bigdeli M, Lai A, Schauer JJ. Seasonal trends, chemical speciation and source apportionment of fine PM in Tehran. Atmospheric Environment. 2017;153:70-82.

20. Shahbazi H, Reyhanian M, Hosseini V, Afshin H. The relative contributions of mobile sources to air pollutant emissions in Tehran, Iran: an emission inventory approach. Emission control science and technology. 2016;2:44-56.

21. Atash F. The deterioration of urban environments in developing countries: Mitigating the air pollution crisis in Tehran, Iran. Cities. 2007;24(6):399-409.

22. Chu Y, Liu Y, Li X, Liu Z, Lu H, Lu Y, et al. A review on predicting ground $PM_{2.5}$ concentration using satellite aerosol optical depth. Atmosphere. 2016;7(10):129.

23. Li X, Zhang X. Predicting ground-level PM2. 5 concentrations in the Beijing-Tianjin-Hebei region: A hybrid remote sensing and machine learning approach. Environmental pollution. 2019;249:735-49.

24. Li T, Shen H, Yuan Q, Zhang L. Geographically and temporally weighted neural networks for satellite-based mapping of ground-level $PM_{2.5}$. ISPRS Journal of Photogrammetry and Remote Sensing. 2020;167:178-88.

25. Huang K, Xiao Q, Meng X, Geng G, Wang Y, Lyapustin A, et al. Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain. Environmental pollution. 2018;242:675-83.

26. He Z, Guo Q, Wang Z, Li X. Prediction of monthly $PM_{2.5}$ concentration in Liaocheng in China employing artificial neural network. Atmosphere. 2022;13(8):1221.

27. Guo B, Wang X, Pei L, Su Y, Zhang D, Wang Y. Identifying the spatiotemporal dynamic of $PM_{2.5}$ concentrations at multiple scales using geographically and temporally weighted regression model across China during 2015–2018. Science of The Total Environment. 2021;751:141765.

28. Chen B, Song Z, Pan F, Huang Y. Obtaining vertical distribution of $PM_{2.5}$ from CALIOP data and machine learning algorithms. Science of The Total Environment. 2022;805:150338.

29. Zhang Y, Sun Q, Liu J, Petrosian O. Long-Term Forecasting of Air Pollution Particulate Matter ($PM_{2.5}$) and Analysis of Influencing Factors. Sustainability. 2023;16(1):19.

30. Ma J, Yu Z, Qu Y, Xu J, Cao Y. Application of the XGBoost machine learning method in PM2. 5 prediction: a case study of Shanghai. Aerosol and Air Quality Research. 2020;20(1):128-38.

31. Zamani Joharestani M, Cao C, Ni X, Bashir B, Talebiesfandarani S. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data.

Atmosphere. 2019;10(7):373.

32. Xiao F, Yang M, Fan H, Fan G, Al-Qaness MA. An improved deep learning model for predicting daily $PM_{2.5}$ concentration. Scientific reports. 2020;10(1):20988.

33. Ho C-H, Park I, Kim J, Lee J-B. $PM_{2.5}$ forecast in korea using the long short-term memory (LSTM) model. Asia-Pacific Journal of Atmospheric Sciences. 2023;59(5):563-76.